DOCUMENT RESUME

ED 322 176 TM 015 321

AUTHOR Mandeville, Garrett K.

TITLE An Alternative Three-Level Model for Estimating SEIs and Their Stability.

PUB DATE Apr 90

NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Academic Achievement; Achievement Gains; Achievement Tests; Educational Trends; *Elementary Schools; *Estimation (Mathematics); Grade 2; Grade 3; Longitudinal Studies; *Models; Primary Education; *Public Schools; *School Effectiveness; State Programs; Testing Programs

IDENTIFIERS Basic Skills Assessment Program; *Educational Indicators; South Carolina

ABSTRACT

After a brief review of recent research concerning school effectiveness, this paper proposes an approach to operationalizing school effectiveness that is based on the achievement of students in a school over a period of time. This three-level model does not follow a single cohort of students as they progress through school. Rather, it uses the achievement results of five consecutive second- and third-grade classes in a given school. The approach uses first stage modeling to obtain school posttest performance indices adjusted for pretest for each of the 5 years. In a second stage, the trend across the 5-year time span is modeled; and in a third stage, the trend parameter estimates are related to school level variables. Each grade level and subject area (reading and mathematics) are considered separately, so the school effectiveness indices (SEIs) do not produce a comprehensive index of school effectiveness. Matched student-level data files for grades two and three in South Carolina for each of the 5 school years in the span from 1984-85 to 1988-89 were used. The complete second-grade data set contained over 194,000 students whose means on the Basic Skills Assessment Program Reading Test (BSAPR) and on the Basic Skills Assessment Program Mathematics Test (BSAPM) were 832 and 810, respectively. The complete third-grade data set contained over 190,000 students whose means on the BSAPR and the BSAPM were 800 and 806, respectively. Data for 148 second-grade and 146 third-grade samples of schools are provided. Results are discussed in terms of the effects of centering, low reliabilities of slopes, and correlates of status and slope. Five data tables are included. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

An Alternative Three-Level Model for Estimating SEIs and Their Stability

ED322176

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Garrett K. Mandeville

University of South Carolina

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GARRETT K. MANDEVILLE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running Head: A Three-Level Model for Estimating SEIs

One of the presentations in the symposium entitled
Assessing School Effectiveness:
New Directions and "Real-World" Implications
at the annual convention of the
American Educational Research Association
Boston, April 1990

TM015321

BEST COPY AVAILABLE

An Alternative Three-Level Model for Estimating SEIs and Their Stability

The quest for assessing the effectiveness of schools to foster learning in the children who attend them has been going on for some time. Although the lines of demarcation are not always clear, there appear to be three somewhat distinct research thrusts in the general area of school effectiveness. The three proposed types of research are: 1) "school effects" (SE) research (e.g., Coleman, et al., 1966); 2) "effective schools" (ES) research (e.g., Edmonds, 1979); and 3) research related to school reward programs (SRPs) (e.g., Wynne, 1984).

As outlined earlier by Kijai, ES research has produced a list of indicators which have been generally embraced by the educational community as valid, i.e, as related to school effectiveness. Thus, a number of school improvement programs have been designed around these indicators and implemented across the nation. Questions do remain, however, regarding the relationships of these indicators to measures of school effectiveness and, furthermore, whether ES variables might function as useful predictors in SE and SRP research efforts.

These questions lead naturally to the issue of how to operationalize school effectiveness. Traditionally monitoring of school improvement and/or attainment has involved some form of regression analysis using either cross-sectional or longitudinally matched data sets (Dyer, Linn & Patton, 1969). Recently, however, the appropriateness of these approaches has been questioned (e.g., Aitkin & Longford, 1986). Most of the current developmental work in computing school effectiveness indices (SEIs) involves operationalizations within the general framework of heirarchical linear models (HLM) (Raudenbush & Bryk, 1986, 1989).

In the first paper, Kijia operationalized school effectiveness using the so-called school gain index (SGI) which is based on student-level regressions and would, therefore, be classified as a "traditional" measure. The SGI is the primary factor in determining which schools receive monetary awards under the South Carolina School Incentive Reward Program (SCSIRP). Since SCSIRP awards are given annually, it is quite logical that the SGI is based on achievement data for a single year only.

On the other hand, one could argue that the identification of schools which have been unusually effective (or ineffective) should be based on a more complete picture of performance than is possible from observing only the most recent achievement results. This would be particularly true for ES researchers who run the risk of making the type of error suggested by analogy in Good and Brophy (1986), namely, of studying a Super Bowl winner a year later when (the 49ers excepted), the odds are against such an exemplary performance.

In the second paper, Kennedy and Mandeville (hereafter, K&M) considered longitudinal data and what appears to be a unique operationalization of effectiveness, one with an emphasis on movement toward equity. They also utilized the HLM approach to modeling so that the SEIs used in that paper would be considered more modern. It is important to note, however, that the SEIs in the K&M paper were based on the performance of a relatively small (in most cases) proportion of the student body in each school and were not aggregated across subject area as is true of the SGIs. Thus, the SGI is clearly a more comprehensive index of school effectiveness than the K&M operationalization.

This paper proposes another approach to operationalizing

school effectiveness which is somewhat similar to the K&M approach in that it is based on the achievement of students in a school over a period of time. Differing from the approach of K&M, however, this study does not follow a single cohort of students as they progress through school. Rather, it utilizes the achievement results of five consecutive 2nd and 3rd grade classes (not to be confused with classrooms) in a given school. The approach here is to use what is referred to as first stage modeling to obtain school posttest performance indices adjusted for pretest for each of the five years. In the second stage, the trend across the five year time span will be modeled and the third stage involves relating the trend parameter estimates to school level variables. Each grade level and subject area (reading and math) will be considered separately, so the SEIs studied here also suffer from the limitation of not producing a comprehensive index of school effectiveness.

Methods

Data Sources

Student Level Test Data

All SC public school children in grades 1-11 are annually tested in the spring with either a CRT or an NRT. For the purpose of this study, only the testing at grades 1-3 is of interest. At these grade levels, the test used is the locally (SC) developed Basic Skills Assessment Program (BSAP) which contains subtests in reading (BSAPR) and mathematics (BSAPM). Scale scores for these tests are horizontally equated (across years) and were used in the analyses both as predictor and criterion variables.

Matched student-level data files for (posttest) grades 2 and 3 for each of the five school years in the span from 1984-85 to 1988-89

were available to the investigator. For the last four years--1985-86 to 1988-89--these files were produced as a byproduct of the data processing for the SCSIRP. As such, the accuracy of the computer matches had been checked visually and edited as necessary. Furthermore, the files for these four years contained records for students who were repeating the grade in question. On the other hand, the 1984-85 data were produced during a feasibility study in which computer matches were not visually edited and repeaters were not included.

Although this indicates that the first data set was somewhat incompatable with the last four, there were a number of reasons which led this investigator to use the data for the full five years. First, it was felt necessary to eliminate data on repeaters throughout the study. The primary reason for this decision was that for repeaters, the pretest was a different test than for nonrepeaters. For example, for 2nd grade repeaters in 1989 the only available pretest was their BSAP2 performance for the Spring of 1988 whereas for promoted students the pretest was their BSAP1 data for the same time period. Although it might have been possible to transform the BSAP2 scores of the repeaters to the scale of the BSAP1 using vertical equating formulas which are available for these lower grades, the validity of applying these formulas to repeaters is open to question. Over the past four years statewide 2nd grade retention rates as estimated from the SCSIRP data have ranged from 4.1% to 5.4% and the corresponding 3rd grade figures are 3.5% to 4.4%.

The second inconsistency was that the 1984-85 data had not been edited. This was not viewed as a serious problem here since the algorithm has been shown to be extremely accurate (Mandeville, 1988)

producing only about .25% mismatches and .8% missed matches without editing. Furthermore, for small schools, where a few aberrant data points would have the most effect on the analyses, error rates can be expected to be lower.

School Level Data

Some obviously important contextual variables to include in school level models were the aggregates of prior achievement which will be referred to as MEANREAD and MEANMATH. These were weighted averages of the pretest means for each school over the five years under investigation. In addition, the percentage of students eligible for free lunch (LUNCHPCT) was obtained as a proxy for SES, again over the five years. As was true of MEANREAD and MEANMATH, LUNCHPCT was calculated separately for 2nd and 3rd graders.

Although the results have been mixed, school size has been shown in some studies to be predictive of achievement. Logically, it should be related to the resources available if one assumes an "economies of scale" argument. Here the average gradesize (GRSIZE), again over the five year period, was estimated by using the total number of 2nd or 3rd graders who were posttested. Also, the total school size was obtained from information on average daily attendance for the 1987-88 school year (ADATOT). Whereas GRSIZE was grade-specific, ADATOT obviously was not.

Two other variables which might logically have to do with school context were obtained. One was an estimate of the percentage of students repeating the grade (REPPCT). This estimate was based on repeater information contained on the SCSIRP data files and was aggregated for two of the five study years, 1986-87 and 1987-88. The other variable was an estimate of the attrition rate at the 2nd or the

3rd grade for the each school. This variable was operationalized in the following way. The number of students who took the posttest but who were not matched (recall that matches included both promoted students and repeaters) were assumed to have been new arrivals to the school district. The attrition rate (ATTRPCT) was computed by expressing this number of students as a percentage of the total posttested. Ramey (personal communication) has suggested what is probably a more defensible operationalization which also takes into account the number of children who were pretested but who left the school during the year. Another problem with the approach used here is that children who move into the system during the school year but leave before the spring testing are not counted.

The final set of school-level variables were those derived from the effective schools survey which has been discussed earlier. For the ES data, the following steps were taken in an attempt to guarantee that the results for each school represented a reasonable proportion of the potential respondents. The editing began by computing ratios of the number of teacher, parent, and student responses to total student enrollment as measured by ADATOT. The distributions of these ratios were inspected and the following--admittedly somewhat arbitrary--decisions were made. Since all teachers in a school were requested to complete the survey, at least one teacher response for each 25 students in ADA was required. For parent data, which was conducted on a sample basis, a ratio of one parent response for each ten students was required. And finally, this same one-to-ten ratio was set as the minimum standard for student data.

It should be recalled that students below grade 4 were not sampled so that it is not surprising that a large number of the

schools considered in this study did not achieve the student standard. Probably a more appropriate reason to discount the student data in this study was the fact that none of the target students--2nd and 3rd graders--were surveyed.

As regards schools lost because of these restrictions, it should be pointed out that of the approximately 550 SC schools with grades 2 and 3, ES data were only available for about 350. Of these schools, over 290 passed both the teacher and parent data edit checks whereas only about 80 passed the student edits. Therefore, student data on the six ES variables was not included. At this stage, then, the sample sizes were 293 schools with 2nd grade and 294 schools with 3rd grade. These variables are denoted as follows: instructional leadership (LEAD_T for teachers and LEAD_P for parents); academic emphasis (ACAD_T and ACAD_P); high expectations (EXPECT_T and EXPECT_P); positive school climate (CLIMAT_T and CLIMAT_P); frequent monitoring (MONITO_T and MONITO_P); and home-school relations (RELAT_T and RELAT_P).

Assumed Constancy of School-Level Data

In the 3rd stage modeling (described below), an implicit assumption is that the school variables represent global constructs which were relatively constant over the five year study period. For many of the variables included, this assumption could have been addressed since data for each year were available. This would be the case, for example, for the school averages which were derived from the student level test records such as MEANREAD, MEANMATH, GRSIZE and LUNCHPCT. For the others, however, the data represented the condition at one (ADATOT, the ES variables) or two (REPPCT and ATTRPCT) points in time. While the assumption of relative constancy was likely largely

satisfied for variables related to achievement, SES, and the like, it is not clear that this would be the case for the ES variables. Especially those focussed on the principal, such as instructional leadership, would seem to be highly sensitive to staffing changes. Also one would expect statewide school improvement efforts to have an effect on the ES variables over time. It is also true that changes in school policy regarding student promotion, changes in school attendance zones, etc., could operate to destroy the veracity of this assumption for most of the other variables included. This is clearly a limitation of the study and points up the need for a more adequate model which would allow changes in these variables over the five year span be related to the dependent variables.

Editing Based on Within-School Sample Size

The other factor which was considered was the number of student records which were processed in the first stage modeling. Variables were created which indicated whether, for each year, the within-school analyses were based on at least 20 records, at least 50 records, or at least 80 records. For the analyses presented here, a minimum of 50 posttest records at that grade-level for each of the five years was required. At grade 2 this produced 148 schools and at grade 3 the n under this restriction was 146. Obviously, these schools do not represent random samples from the populations of schools housing 2nd or 3rd graders.

First Stage Modeling

In first stage modeling both the BSAPR and BSAPM for the prior year were used as predictors for each of the posttest versions of BSAPR and BSAPM and each of the predictors was centered at its mean for the total sample over the five year period. The mean over the

study period was used rather than the means for the individual years so that changes in the performance of the typical student over the five year study period could be assessed. Thus, the following set of equations were fit by OLS for each school for each of the five years.

Grade 2: BSAP2R = f(DBSAP1R,DBSAP1M) and BSAP2M = f(DBSAP1R,DBSAP1M)

Grade 3: BSAP3R = f(DBSAP2R,DBSAP2M) and BSAP3M = f(DBSAP2R,DBSAP2M)

In the above BSAP2R denotes the BSAP reading score for 2nd graders and DBSAP1R and DBSAP1M are the centered 1st grade reading and math scores for these same students. At this stage there was no restriction as to the number of students per school; the limitation to consider only those schools with 50 or more students was introduced later. Depending on which of the five years of data were being processed, between 551 and 560 schools with 2nd grades were involved in this 1st stage modeling and there were between 549 and 562 schools with grade 3.

Because of the large number of schools, it was not possible to insure that both predictors were useful in each case. For that matter, it is likely that for many schools, DBSAPM contributed little to the prediction of BSAPR. The converse is generally not true, however, since DBSAPR is usually an important predictor for BSAPM.

Second and Third Stage Modeling

The intercepts from the first stage analyses--B0R for reading and B0M for math--were used as the dependent variables for the second stage analyses. They represent predicted reading or math achievement for the average (on both predictors) student for each of five years. It is important to recall that the "average student" was defined globally, i.e., the pretest scores were centered to the predictor

means aggregated over the five year period. Since achievement has generally been increasing in South Carolina, the profiles for each school also tend to increase. In stage two analyses, regressions were again run at the school level. The five values of B0R or B0M constitute the dependent variable and year becomes the independent variable. As described in the Bryk and Raudenbush (1987) paper on growth curves, a general polynomial model--such as the one presented as "model (1)" in the K&M paper--could be fit to these data. In practice, one usually inspects the mean profile to assess the degree of the polynomial and this approach was used here. Assuming that the growth curve for math fits a linear model, the equation fit will be:

$$B0M_{jt} = \pi_{0j} + \pi_{1j}A_{jt} + R_{jt}$$

where j is the subscript indexing school, t is the index for time, A_{jt} represents the specific coding for the time factor at time t, and R_{jt} is a random error. In this study, A_{j1} will be coded as 0, A_{j2}, as 1, an so forth. Then the parameter π_{0j} represents the true initial status (at the end of 1985) for school j and π_{1j} is the linear growth parameter, i.e., the average increase per year. If the above model does not suffice, higher order parameters--which reflect departures from linearity--will be needed.

Assuming a linear model, we might consider the merits of the estimates of π_{0j} and π_{1j} as measures of the effectiveness of school j. π_{0j} as defined here is not particularly appealing, in that it estimates historically rather dated effectiveness. The estimate of π_{1j}, on the other hand, assesses the average growth over the five year period and is much more appealing. Although not capturing growth in

securing a more equitable distribution of achievement across SES as in the K&M approach, it is measuring an aspect of effectiveness which should be important to ES researchers.

While the results of estimating the above model, referred to as the "within schools" model, may provide some useful information on school effectiveness, it is important to realize that a number of potentially important factors which affect school achievement have not been accounted for. In fact, the only variables which were included in the first stage modeling were those dealing with the prior achievement of the students. Thus, all schools are considered equal on the other factors which might logically effect school achievement. Some of these factors are beyond the control of the school and are usually referred to as school contextual or school composition variables. Obviously, these should be controlled in some fashion. Others reflect school policies and practices and these should <u>not</u> be controlled if the objective is to measure SEIs. That is, to the extent that these variables promote achievement, schools should not be penalized by removing their effects from SEIs (see, e.g., Raudenbush & Bryk, 1989; Willms & Raudenbush, 1989).

Some of the school-level variables described above are clearly beyond the control of the schools and should be considered context variables. Included here would be the aggregates LUNCHPCT, MEANREAD and MEANMATH. The variables related to school size are obviously related to policy decisions at some level but should probably be considered to be beyond the purview of the teaching and administrative staff at a particular school. Since schools have little control over the community setting, it is likely that ATTRPCT is largely contextual in nature. On the other hand, school policy likely has much to do with

promotion decisions so that REPPCT should be considered in the policy arena. Finally, the objects of the ES variables seem to be related to school policy for, if they were not, they could not be the focus of school improvement efforts.

Third stage modeling provides an opportunity to take into account these school level variables. The estimates of the πs become the dependent variables in the model which was described as "model (2)" in K&M. Such analyses are of obvious import to SE researchers whose objective is to identify and assess the strength of the association between such variables and some operationalization of achievement. For SRP studies, the differentiation between contextual variables and those related to school policy would be important. However, considering the exploratory nature of this investigation, all school level variables described above, without distinction, will be used to fit the 3rd stage or between school model.

Estimation in the 2nd and 3rd models is conducted simultaneously within the HLM framework. Model 2 is first fit by OLS and then the information contained in Model 3--which is fit by Generalized Least Squares (GLS)--is used to adjust the OLS estimates. The analyses will be conducted using the HLM computer program developed by Bryk, Raudenbush, Seltzer and Congdon (1986).

Strategy

In data analysis using HLM, some important advantages accrue. First, an unconditional between school model is fit in which none of the school level variables are included. An important outcome of this is that an estimate of the true parameter variation is obtained and a test of whether it differs significantly from zero is provided. This is important since only true parameter variation can be explained by

the school predictors. In effect, if the hypothesis of zero parameter variance cannot be rejected, the analysis ceases. Assuming nonzero true parameter variance, the school level variables are included in the model and, hopefully, at least some of them will be shown to be related to the dependent variable. If so, the reduction in the parameter variance due to the set of predictors provides important evidence regarding their adequacy in accounting for the variation. Two statistics, both comparable to the multiple R^2, are obtained: one to indicate the percent of true parameter variance accounted for by the predictors and the other to indicate the percent of total observed parameter variance (includes true variance and sampling error) which is explainable in terms of the set of predictors. Hopefully, this somewhat brief discussion will help the reader to follow the results presented below.

Results

The 1st stage regressions were conducted as outlined above. The complete (all five years) 2nd grade data set contained over 194,000 students whose means on the BSAPR and BSAPM were 832 and 810 respectively. For the 3rd graders, over 190,000 were involved and the grand means for reading and math were 800 and 806. For the 1st stage regression analyses, the predictors were centered at these values.

Table 1 contains summary statistics for the 2nd grade (N = 148) and the 3rd grade (N = 146) samples of schools which satisfied the requirement of a minimum of 50 students for each of the five years in the study. Variables that are grade-specific are denoted with a "2/3" appended to the variable name. Since most of the schools in the two samples are the same, the means on the school level predictors are very similar for the two grade levels.

Table 1 about here

The means on B0R and B0M by year suggest that in each case a linear trend will probably be sufficient to fit the data for most of the four grade level-subject area combinations. The major exception may be the grade 2 reading data where improvement seems to have leveled-off in recent years. Nevertheless, the correlation between the coded values of time mentioned above (0, 1, ...) and B0R for grade 2 is still substantial at .965. Although large, this r is smaller than the corresponding figure of .996 for grade 2 math and the two rs for the 3rd grade data, both of which are close to .99. Nevertheless, in each case, fit to the simpler linear model was initially attempted. The checks of the model described in K&M were implemented and, in most cases, the data fit reasonably well. In a few cases the normally of the empirical Bayes residual was questionable, however.

The results for 2nd grade math are reported in Table 2 and will be discussed first. Since the tables are all basically formatted the same, and since each contains an extensive amount of information, a detailed discussion of Table 2 is in order.

Table 2 about here

In the spring of 1985, a typical SC 2nd grader scored about 783 on the BSAP math test. Over the following four years, his or her score increased by about 11 points per year. Obviously initial status will always be significantly different from zero. The fact that the average linear growth (this is the weighted average of the OLS

estimates and, therefore, referred to as a fixed effect) also differs from zero is evidence that math achievement has been on the rise in SC over these five years.

The next item of importance in the rather comprehensive table is the information on the true variance in the parameter estimates which is in the second portion of the table (Random Effects). For this portion of the table the estimates of the π_{0j} and π_{1j} are considered as random variables. Here, the chi square test indicates that the true variance for each of these parameters is significantly different from zero, indicating that the analysis may proceed. The reliabilities of these two "measurements", is reported in the third section of the table, headed Model Statistics. These are the ratios of the estimated true variance to the total observed parameter variance. Here the estimate of initial status has a rather high reliability (.741) whereas the linear growth parameter is less reliable with a value of .426. It is not surprising that the latter have a smaller proportion of true variability since regression coefficients generally contain considerably more sampling variance than do intercepts which function more like sample means.

What of our school level variables were related to the estimates of π_{0j} and π_{1j} is the next question to consider. For initial status (π_{0j}), four variables were statistically significant at the .05 level. One might anticipate that the general level of achievement of the school as indexed by MEANREAD and MEANMATH would have an effect on initial status, even though pretest achievement was controlled in the 1st stage regression analyses. Indeed, this is the case here but the interpretation does not appear to be a simple one. MEANREAD has a positive coefficient while the b for MEANMATH is negative. Since

MEANREAD and MEANMATH are positively correlated (r = .86) and both have modest positive correlations (in the .2 to .3 range) with initial status, MEANMATH appears to be functioning as a suppressor for MEANREAD. The reason for this phenomena is not clear.

REPPCT is also a significant predictor of initial status with a negative coefficient. This indicates that schools retaining larger percents of their 2nd graders were lower on initial status. The most straightforward explanation for this relationship is that schools who retain a larger portion of their students were failing to provide adequate learning situations. (Parenthetically, the leaders of such schools are also apparently unfamiliar with the prepondence of research which indicates that retention is ineffective.) Finally, instructional leadership as perceived by teachers (LEAD_T) is also a significant predictor of initial achievement, but with an unexpected negative coefficient. Since both REPPCT and LEAD_T were also significant predictors of the growth parameter, they will be discussed further below.

Another question to ask is the extent to which the predictors included have explained the variance in initial status. Information on this issue is included in the third portion of the table under the two headings: "Percent of total variance explained" and "Percent of parameter variance explained." The former, denoted as R^2, is 16.8% for initial status and the latter R^{2*} is 22.7. Thus, the four school level variables have accounted for somewhat less that one-fourth of the predictable variance in initial status. Obviously a number of other variables which were not available to this investigator would be needed to provide a more definitive model. The last section in the second portion of the table, "Residual variance.." indicates that the

remaining variance is significantly different from zero. An objective of the HLM modeling approach is to identify and include other relevant predictors until the residual variance does not statistically differ from zero.

As indicated above, it is not surprising that the estimates of π_{1j} are less reliable than those of the π_{oj}. However, the results on the significance of the variance of the growth estimates indicate that true variance in these slopes does exist. Since growth would appear to be more important than initial status, significant predictors of growth are of considerable substantive interest. There are three listed--GRSIZE, REPPCT, and LEAD_T--and all have positive coefficients. Larger growth parameter estimates appear to occur in larger schools (at least those with more 2nd graders), those which retain a larger percentage of their 2nd graders, and those where the teachers perceived higher levels of leadership by the principal. One possible explanation of the significance of REPPCT is an artifact of the design used here, namely the exclusion of repeaters from the analysis. Since repeaters almost by definition are low achievers, excluding them has an artificial positive effect on schools with large retention rates.

A more substantive explanation is also possible. It takes into account the fact that schools which retain large portions of their students probably were among the least successful in 1985. The negative coefficient for REPPCT as a predictor of initial status supports this claim. Those schools with the most room for improvement should be expected (hopefully) to produce the largest yearly gains and this is the case here. Evidence for this is on the last line of the table where the correlation between initial status and linear growth

Table 3 about here

points vs. 11 at grade 2. Since the two scales are not directly comparable, however, it is impossible to compare the two and say that the rate of math achievement gain was greater in grade 3 than in grade 2. Once again, true parameter variance differs from zero for both initial status and linear growth, but the proportions of total variance which were estimated to be true are much smaller, leading to reliabilities of only about .47 for initial status and .29 for growth. In retrospect, a possible explanation for this is that a 2nd degree polynomial is needed to fit the growth curve for grade 3 math to account for the unusual jump in performance between 1988 and 1989, i.e., a 22.5 point gain vs. about a 15 point average gain for the earlier years (see Table 1).

Taken at face value, the 3rd grade math results in Table 3 are somewhat disappointing. The reader will note that two variables which were not statistically significant--ADA and GRSIZE--are included in the table. The only significant predictor of initial status was REPPCT, a result consonant with the grade 2 math results. Also, only one predictor of growth was significant, with a positive coefficient, and again having to do with perceptions of the principal's instructional leadership. The difference is that here parental perceptions (LEAD_P) was significant whereas for grade 2 teacher perceptions were involved. On this data set these two variables (LEAD_T and LEAD_P) are positively correlated--as one would expect and hope--but the r is only about .40. In spite of this difference, it is again encouraging that an index of leadership has an impact on growth

in math achievement. The 3rd grade math results indicate a much lower (absolute) association between initial status and growth rate (-.297) than was found for 2nd grade math. The other curious feature of these results in comparison to those for grade 2 math is the fact that--in the sense of the percent of true variance explained--the predictors were more sensitive to variation in slope than initial status (20.5% vs. 10.9%). If as suggested above the growth parameter is the more important of the two, the inability to predict initial status more accurately may be of lessor import, however.

Let us next turn to 3rd grade reading achievement reported in Table 4. We first note the modest reliabilities associated with the

---

Table 4 about here

---

two measures; the .487 for initial status and .232 for slope are similar in magnitude to the figures for grade 3 math. Second, although a number of predictors are listed for growth rate, none are statistically significant at traditional levels. Thus, the report in the table that 17.1% of the true variation in slope is accounted for is questionable: it would have been more appropriate to eliminate these four predictors and estimate the revised model. As far as initial status is concerned, a number of variables were retained and most of them (ADATOT and EXPECT_T being the exceptions) at least approach the .05 significance level. The results are not amenable to simple interpretation, however. MEANREAD and MEANMATH appear with different signs as was true for grade 2 math; although unappealing, the suppressor idea presented earlier may apply here also. GRSIZE is the only other significant predictor and has a positive coefficient.

ATTRPCT and RELAT_T approach significance and also have positive regression coefficients.

Let us finally discuss the results for grade 2 reading which are reported in Table 5. When the linear model was initially fit to these

---

Table 5 about here

---

data, the reliabilities were quite low and, since a curvilinear trend was suggested by inspection of the means, a quadratic model was fit. Since the reliabilities of the three parameter estimates were reasonably large, this seemed to be a more adequate model for the data. Unfortunately none of the predictors had sufficient explanatory power to be judged significant for either the linear or quadratic terms. Thus, the only finding of substance is that schools with a larger percentage of low SES students, as indicated by larger values of LUNCHPCT, tended to have lower values on initial status.

Discussion

The two objectives of this paper were rather modest. One was to consider the characteristics of two operationalizations of school effectiveness: one which focussed on achievement at a given point in time (initial status was chosen here) and the other having to do with growth in achievement over a period of time. Because of the coding of time used here, the intercept captured the achievement level as of 1985 (initial status) and, consequently would only be of historical interest to researchers attempting to identify effective schools (ES) or those charged with making school reward decisions (SRP). However, different codings could have been employed which would have remedied this situation so we should not discount the importance of estimation

of the intercept in this growth formulation. For example, if time had been centered at the middle of the five year period (coding of -2, -1, 0, 1, 2) the intercept would have represented the average performance of the typical student in the Spring of 1987 and centering on the final year (-4, -3, ..., 0) would have produced an intercept which represented average performance in the Spring of 1989. The latter, being the most current, would probably be of most interest.

The Effects of Centering

It is important to realize, however, that the results presented here--where the intercept represented initial (end of 1985) status--would not necessarily prevail under these different codings. In fact, when the 2nd grade math data were rerun with the above codings, a number of changes occurred. When time was centered at 1987 the (unconditional) reliability of the intercept estimates increased from the .741 reported here to .809. This is quite logical since the sampling error of estimates near the mean of the predictor is smaller than at the extremes. More surprising, however, was the fact that when the data were centered at the most recent year (Spring 1989) the reliability of the estimates dropped to .487. Thus, for the coding which would render the intercept most useful, less than half of its variance is true variance. Since the slope is not affected by the centering, the results for the unconditional model regarding slope were the same as reported above.

When the between models reported above were rerun for these alternative centerings, the results were very interesting. MEANREAD, MEANMATH, and LUNCHPCT remained as significant predictors of the intercept in both cases but LEAD_T did not. In fact, when time was centered at 1989, the (nonsignificant) sign of the coefficient for

LEAD_T was positive. This finding supports the earlier speculation offered to explain the negative and significant coefficient for LEAD_T under the "initial status" centering. The between school model results for linear growth also changed when time was centered at 1989 in that GRSIZE was not deemed to be a significant predictor of slope. Although this issue will not be addressed further here, it points up that fact that because of the simultaneous nature of the estimation process, the decision regarding where to center the data is crucial. This means that, although the slope parameters were of primary interest here, the results presented in reference to them are not independent of the centering decision.

The SEIs based on linear growth, it was claimed, should be of interest to ES researchers but whether this approach is reasonable for SRP research is more debatable. The issue of the stability of SEIs developed according to this formulation was only treated tangentially, via the true variance exhibited by these growth parameter estimates. In each of the four (grade level-subject area) cases considered, statistical tests indicated that significant true variation in these potential SEIs did exist. The "reliabilities" of these measurements, however, tended to be small, being greater than .4 only in the two 2nd grade analyses. This means that a large portion of the increasing trend observed <u>cannot</u> be explained, at least not within the framework of a linear trend. Initial status, on the other hand, was always more reliable than trend, with values larger than .7 for the two 2nd grade analyses.

<u>Low Reliabilities of Slopes</u>

There are at least two explanations for the low reliabilities associated with the estimated slopes. First, the general trend may not

be linear and we saw some evidence to support this explanation for 2nd grade reading and possibly 3rd grade math. The other explanation--which most likely accounts for the majority of the sampling error--is that at the individual school level considerable deviation from linearity occurs.

This possibility was briefly explored by computing the correlations between year and B0R and year and B0M for the 2nd grade sample of 148 schools. For reading, the median r was .82 and the interior 80 percent of the distribution ranged from P_{10} = .57 to P_{90} = .96. There were, however, four negative correlations between -.38 and -.23 and 5% of the correlations were less than .28. The largest r was .997. For math there was even more variation. The median r was .81 and the interior 80 percent of the rs ranged from .21 to .96. The 5th percentile for math was an r of .05, and five schools produced negative correlations. While four of these negative rs were small in magnitude, one was -.77. It was interesting to note that the school with the -.77 r in math also had a negative r in reading and that one other school produced negative rs for both subjects. Something rather remarkable must have occurred in these two schools during this time span.

Obviously, there is considerable variation in the 2nd grade school profiles--at least as far as deviation from linearity is concerned--in both subjects, but apparently more in math than reading. Thus, it is curious that the reliability of the estimated slopes when a linear model was attempted for the 2nd grade reading data (.215) was lower than for 2nd grade math (.426), so low in fact that a quadratic model was chosen which produced the reliability of .429 reported earlier.

Correlates of Status and Slope

The second objective was to determine whether the various school level variables included in the model were useful in the prediction of the growth parameters. The inclusion of such variables in the construction of SEIs is important for a number of reasons. For example, school level modeling is the only sensible way to include school level variables which are not based on student aggregation such as teacher characteristics (e.g., average salary, experience) or school characteristics (e.g., size, organizational structure, principal variables). In addition, it is well-known that student background variables often produce compositional (school level) effects beyond what is captured by modeling at the student level.

For the purpose of identifying effective schools or making awards, therefore, it is important to use the information on these variables. The logical way to proceed would be to use only the contextual variables for school level modeling and consider the deviations of the actual outcomes from the school level predictions as the SEIs. Raudenbush and Bryk (1989), however, have shown that a more appropriate--probably necessary--strategy is to include both contextual and policy variables in the analysis. The reason is that there is likely to be an association between context and policy variables and unless both are included, the removal of the effect of context variables will also remove a portion of the effect associated with policy factors. They suggest that the association is likely to be positive, i.e., that schools advantaged by context will also have policies which promote achievement. For these reasons, it was considered important here to investigate the utility of the assembled school predictors.

As suggested above, some form of residual between the estimates based on the school data alone (here the OLS initial status and slope from the 2nd stage) and the predictions based on the school level model (3rd stage GLS) are used to assess the school effect. The HLM solution, however, is to offer an alternative to the 2nd stage OLS estimates. These empirical Bayes (EB) estimates are weighted averages of the 2nd stage OLS estimates and the 3rd stage model predictions where the weights are related to the reliabilities mentioned above. In essence, the 2nd stage estimates are "shrunk" toward the 3rd stage predictions if the former are deemed to be of low reliability (have large sampling error) or remain close to the OLS values if they are deemed to be highly reliable. The resulting EB estimates have been shown to have smaller mean square error than the OLS estimates.

On the substantive level, the school level variables considered here demonstrated limited power to predict the 2nd stage results. In all but one case, 20-25% of the true variance in initial status was predictable; for slope, the figures ran from 16-20% for the three cases where significant predictors were identified. Furthermore, the inclusion of predictors of questionable merit suggests that these values are overestimates.

In three of the cases the most important predictors of initial status were contextual variables which were aggregates of student SES or achievement such as MEANREAD, MEANMATH, and LUNCHPCT. REPPCT also was seen to be an important predictor of initial status but, as noted earlier, this may well be an artifact of the design in which repeaters were eliminated so a substantive interpretation does not seem in order. The only ES variable which was significant beyond the .05 level for initial status was instructional leadership with a

negative coefficient for grade 2 math.

Significant predictors of linear growth were found in two cases, 2nd and 3rd grade math. In both cases variables having to do with the principal's leadership skills entered with positive weights. The size of the 2nd grade class was significant and school size as measured by ADATOT approached significance for grade 3, both coefficients being positive.

Closing Comments

It was hoped that this study would provide information on a method of assessing school effectiveness which was based on the performance of primary school students over five successive school years. Since the actual mechanics of producing SEIs was never dealt with directly and the stability issue was treated in only a limited fashion, the paper has fallen far short of what one might expect considering its title. Indeed, this paper was exploratory in nature.

As noted earlier, the estimated linear growth parameter, adjusted for school context, might be a useful approach to identifying exceptional schools in the ES sense. If only the two most current posttest years were used, the results might also be useful for SRP studies. However, it still seems as though it would be necessary to conduct the analyses separately for each grade level and, unless a composite reading and math score were used, for each subject area. For SRP applications, these separate SEIs would have to be aggregated. Though less elegant than analyzing the data for all public school children in the state (regardless of grade level) in the context of a comprehensive model which posits a general school effect which functions across grades, it is the only "modern" approach which seems remotely practical at this time.

The thorny issue of the distribution of achievement as a function of ability was not dealt with in this paper. Rather, the results here referred to the student of "average ability" over the extended five year period. This approach begged the issue of the dimensionality of the school effects (see Raudenbush & Bryk, 1989). The limited evidence on this issue, however, suggests that (because the slopes are so unreliable) when the EB residuals are used, school effects are effectively constant as a function of previous achievement (see Tate, Im, & Piotrowski, 1989).

References

Aitkin, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. Journal of the Royal Statistical Society, Series A, 149, 1-42.

Bryk, A. S., Raudenbush, S. W., Seltzer, M. & Congdon, R. T. (1986). An introduction to HLM: Computer program and user's guide. University of Chicago, Department of Education.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of heirarchical linear models to assessing change. Psychological Bulletin, 101, 147-158.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). On Equality of Educational Opportunity. Washington, DC: US Department of Health, Education and Welfare.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy scores based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Edmonds, R. R. (1979). Effective schools for the urban poor. Educational Leadership, 37(1), 15-18, 20-24.

Good, T. L., & Brophy, J. E. (1986). School effects. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 570-604). New York: Macmillan.

Mandeville, G. K. (1988, April). A record linking system for research on school improvement. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 301 167)

Raudenbush, S. & Bryk, A. S. (1986). A heirarchical model for studying school effects. Sociology of Education, 59, 1-17.

Raudenbush, S. W. & Bryk, A. S. (1989). Quantitative models for estimating teacher and school effectiveness. In R. D. Bock (Ed.), Multilevel analysis of educational data. San Diego: Academic Press.

Tate, R. L., Im, S., & Piotrowski, W. (1989, March). School comparisons with heirarchical linear models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Willms, J. D. & Raudenbush, S. W. (1989). A longitudinal heirarchical linear model for estimating school effects and their stability. Journal of Educational Measurement, 26, 209-232.

Wynne, E. A. (1984). School award programs: Evaluation as a component of incentive systems. Educational Evaluation and Policy Analysis, 6, 85-93.

Table 1
Summary Statistics for 2nd and 3rd Grade Samples
With a Minimum of 50 Student Records Per Year

	Grade 2 (N = 148)		Grade 3 (N = 146)	
Variable	Mean	SD	Mean	SD
Potential Predictors				
General				
MEANLCH-2/3	33.83	18.27	33.84	18.37
MEANREAD-2/3	812.19	18.30	806.59	24.88
MEANMATH-2/3	837.56	25.26	803.78	25.17
GRSIZE-2/3	94.39	36.79	93.11	33.82
ADATOT	523.17	145.36	530.75	146.63
REPPCT-2/3	4.35	3.65	3.72	3.50
ATTRPCT-2/3	11.51	5.08	11.38	5.24
Effective Schools				
LEAD_T	4.30	0.29	4.29	0.29
ACAD_T	4.27	0.17	4.26	0.18
EXPECT_T	4.22	0.18	4.21	0.18
CLIMAT_T	4.18	0.30	4.15	0.30
MONITO_T	4.12	0.21	4.11	0.21
RELAT_T	4.35	0.27	4.35	0.26
LEAD_P	3.81	0.22	3.81	0.20
ACAD_P	4.09	0.13	4.08	0.13
EXPECT_P	3.96	0.15	3.95	0.14
CLIMAT_P	3.92	0.18	3.90	0.18
MONITO_P	3.82	0.14	3.82	0.13
RELAT_P	4.10	0.15	4.09	0.14
Dependent Variables				
Math Intercepts				
B0M85	782.99	28.97	769.13	26.64
B0M86	793.94	27.01	774.40	31.94
B0M87	807.50	25.26	797.73	32.67
B0M88	814.53	19.18	813.93	30.56
B0M89	829.41	23.02	836.46	36.00
Reading Intercepts				
B0R85	773.06	23.85	777.87	17.48
B0R86	788.61	21.67	790.95	18.39
B0R87	807.38	21.88	793.71	15.77
B0R88	815.99	18.04	806.40	18.28
B0R89	818.80	19.94	816.71	18.61

Table 2
Summary of HLM Results for the
Effects of Various School Variables on
Growth Parameters for Grade 2 Math

Effects	Coeff	SE	z	p	Estimate	Chi Sq	df	p
Fixed Effects								
Average within-school eqn.								
Initial status	782.987	2.209	354.44	.000				
Linear growth	11.343	.606	18.73	.000				
Effect on Initial Status								
MEANREAD	.512	.129	3.97	.000				
MEANMATH	-.234	.096	-2.44	.015				
REPPCT	-2.626	.571	-4.60	.000				
LEAD_T	-15.712	6.915	-2.27	.023				
Effect on Linear Growth								
GRSIZE	.024	.010	2.48	.013				
REPPCT	.323	.161	2.01	.045				
LEAD_T	4.822	2.004	2.41	.016				
Random Effects								
Total (observed) variance in parameter estimates								
Initial Status, var(π_{0j})					722.27			
Linear growth, var(π_{1j})					54.29			
Uncond'l (true) variance in parameter estimates								
Initial Status, var(π_{0j})					535.25	567.73	147	.000
Linear growth, var(π_{1j})					23.12	256.03	147	.000
Residual variance in growth parameter estimates								
Initial Status, var(π_{0j})					413.82	465.87	143	.000
Linear growth, var(π_{1j})					19.35	233.35	144	.000
Model Statistics								
ML Estimate of σ^2 (unconditional model)					311.69			
ML Estimate of σ^2 (conditional model)					310.38			
Reliability (Unconditional Model)								
Initial Status, var(π_{0j})					.741			
Linear growth, var(π_{1j})					.426			
Percent of total variance explained (R^2)								
Initial Status, var(π_{0j})					16.8			
Linear growth, var(π_{1j})					6.9			
Percent of parameter variance explained (R^{2*})								
Initial Status, var(π_{0j})					22.7			
Linear growth, var(π_{1j})					16.3			
Correlation of initial status and linear growth (unconditional model)					-.817			

Table 3
Summary of HLM Results for the
Effects of Various School Variables on
Growth Parameters for Grade 3 Math

Effects	Coeff	SE	z	p	Estimate	Chi Sq	df	p
Fixed Effects								
Average within-school eqn.								
Initial status	763.491	2.226	343.01	.000				
Linear growth	17.420	.782	22.27	.000				
Effect on Initial Status								
GRSIZE	.095	.055	1.74	.082				
ADATOT	-.019	.015	-1.21	.227				
REPPCT	-1.999	.482	-4.15	.000				
Effect on Linear Growth								
ADA	.009	.005	1.89	.059				
LEAD_P	14.817	5.482	2.70	.007				
RELAT_P	-13.570	7.599	-1.79	.074				
Random Effects								
Total (observed) variance in parameter estimates								
Initial Status, var(π_{0j})					723.34			
Linear growth, var(π_{1j})					89.36			
Uncond'l (true) variance in parameter estimates								
Initial Status, var(π_{0j})					341.91	274.97	145	.000
Linear growth, var(π_{1j})					25.78	203.81	145	.001
Residual variance in growth parameter estimates								
Initial Status, var(π_{0j})					304.53	257.41	142	.000
Linear growth, var(π_{1j})					20.50	188.77	142	.005
Model Statistics								
ML Estimate of σ^2 (unconditional model)					635.73			
ML Estimate of σ^2 (conditional model)					635.47			
Reliability (Unconditional Model)								
Initial Status, var(π_{0j})					.473			
Linear growth, var(π_{1j})					.289			
Percent of total variance explained (R^2)								
Initial Status, var(π_{0j})					5.2			
Linear growth, var(π_{1j})					5.9			
Percent of parameter variance explained (R^{2*})								
Initial Status, var(π_{0j})					10.9			
Linear growth, var(π_{1j})					20.5			
Correlation of initial status and linear growth (unconditional model)					-.297			

Table 4
Summary of HLM Results for the
Effects of Various School Variables on
Growth Parameters for Grade 3 Reading

Effects	Coeff	SE	z	p	Estimate	Chi Sq	df	p
Fixed Effects								
Average within-school eqn.								
Initial status	778.503	1.284	606.50	.000				
Linear growth	9.313	.428	21.75	.000				
Effect on Initial Status								
MEANREAD	-.163	.065	-2.51	.012				
MEANMATH	.226	.072	3.14	.002				
GRSIZE	.065	.031	2.08	.037				
ADATOT	-.015	.010	-1.47	.141				
ATTRPCT	.495	.262	1.89	.059				
EXPECT_T	-4.613	6.934	-.67	.506				
RELATE_T	10.424	5.477	1.90	.057				
Effect on Linear Growth								
ADATOT	.005	.003	1.60	.101				
ATTRPCT	-.111	.090	-1.22	.222				
LEAD_P	1.742	2.139	.81	.416				
MONITO_P	-3.737	3.020	-1.24	.216				
Random Effects								
Total (observed) variance in parameter estimates								
Initial Status, var(π_{0j})					240.55			
Linear growth, var(π_{1j})					26.76			
Uncond'l (true) variance in parameter estimates								
Initial Status, var(π_{0j})					117.17	282.71	145	.000
Linear growth, var(π_{1j})					6.20	188.69	145	.009
Residual variance in parameter estimates								
Initial Status, var(π_{0j})					86.87	242.91	138	.000
Linear growth, var(π_{1j})					5.14	179.43	141	.016
Model Statistics								
ML Estimate of σ^2 (unconditional model)					205.63			
ML Estimate of σ^2 (conditional model)					206.48			
Reliability (Unconditional Model)								
Initial Status, var(π_{0j})					.487			
Linear growth, var(π_{1j})					.232			
Percent of total variance explained (R^2)								
Initial Status, var(π_{0j})					12.6			
Linear growth, var(π_{1j})					4.0			
Percent of parameter variance explained (R^{2*})								
Initial Status, var(π_{0j})					25.9			
Linear growth, var(π_{1j})					17.1			
Correlation of initial status and linear growth (unconditional model)					-.386			

Table 5
Summary of HLM Results for the
Effects of Various School Variables on
Growth Parameters for Grade 2 Reading

Effects	Coeff	SE	z	p	Estimate	Chi Sq	df	p
Fixed Effects								
Average within-school eqn.								
Initial status	771.904	1.958	394.21	.000				
Linear growth	22.066	1.634	13.51	.000				
Quadratic	-2.545	.371	-6.87	.000				
Effect on Initial Status								
LUNCHPCT	-.440	.077	-5.73	.000				
MEANREAD	-.100	.072	-1.38	.168				
REPPCT	-.619	.353	-1.76	.079				
EXPECT_P	13.245	7.747	1.71	.087				
Effect on Linear Growth								
No Variables								
Effect on Quadratic term								
No Variables								
Random Effects								
Total (observed) variance in parameter estimates								
Initial Status, var(π_{0j})					567.45			
Linear growth, var(π_{1j})					395.11			
Quadratic					20.34			
Uncond'l (true) variance in parameter estimates								
Initial Status, var(π_{0j})					406.58	518.53	147	.000
Linear growth, var(π_{1j})					169.37	257.29	147	.000
Quadratic					7.37	230.52	147	.000
Residual variance in parameter estimates								
Initial Status, var(π_{0j})					302.09	419.46	143	.000
Linear growth, var(π_{1j})					NA			
Quadratic term					NA			
Model Statistics								
ML Estimate of σ^2 (unconditional model)					181.63			
ML Estimate of σ^2 (conditional model)					181.45			
Reliability (Unconditional Model)								
Initial Status, var(π_{0j})					.717			
Linear growth, var(π_{1j})					.429			
Quadratic					.362			
Percent of total variance explained (R^2)								
Initial Status, var(π_{0j})					18.4			
Linear growth, var(π_{1j})					NA			
Quadratic term					NA			
Percent of parameter variance explained (R^{2*})								
Initial Status, var(π_{0j})					25.7			
Linear growth, var(π_{1j})					NA			
Quadratic term					NA			
Correlation of initial status and Linear growth (unconditional model)					-.591			
Quadratic term (unconditional model)					.480			